

# Applied Regression Analysis

THIRD EDITION

Norman R. Draper

Harry Smith

WILEY SERIES IN PROBABILITY AND STATISTICS  
TEXTS AND REFERENCES SECTION

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *Vic Barnett, Ralph A. Bradley, Noel A. C. Cressie, Nicholas I. Fisher,  
Ian M. Johnston, J. B. Kadane, David G. Kendall, David W. Scott,  
Bernard W. Silverman, Adrian P. M. Smith, Josef L. Teugels, Geoffrey S. Watson,  
J. Stuart Hunter, Emeritus*

A complete list of the titles in this series appears at the end of this volume.



# Contents

Preface	xiii
About the Software	xvii
0 Basic Prerequisite Knowledge	1
0.1 Distributions: Normal, $t$ , and $F$ , 1	
0.2 Confidence Intervals (or Bands) and $F$ -Tests, 4	
0.3 Elements of Matrix Algebra, 6	
1 Fitting a Straight Line by Least Squares	15
1.0 Introduction: The Need for Statistical Analysis, 15	
1.1 Straight Line Relationship Between Two Variables, 18	
1.2 Linear Regression: Fitting a Straight Line by Least Squares, 20	
1.3 The Analysis of Variance, 28	
1.4 Confidence Intervals and Tests for $\beta_0$ and $\beta_1$ , 34	
1.5 $F$ -Test for Significance of Regression, 38	
1.6 The Correlation Between $X$ and $Y$ , 40	
1.7 Summary of the Straight Line Fit Computations, 44	
1.8 Historical Remarks, 45	
Appendix 1A Steam Plant Data, 46	
Exercises are in "Exercises for Chapters 1-3", 96	
2 Checking the Straight Line Fit	47
2.1 Lack of Fit and Pure Error, 47	
2.2 Testing Homogeneity of Pure Error, 56	

This book is printed on acid-free paper. ©

Copyright © 1998 by John Wiley & Sons, Inc. All rights reserved.

Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923. For those organizations that have been granted a photocopy licence by CCC, a separate system of payment has been arranged. The fee code for users of the Transactional Reporting Service is 0883-8502/98 \$05.00. This publication is registered at the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923. Printed in the United States of America. 0012 (212) 850-6011, fax (212) 850-6008, E-Mail: PERMREQ@WILEY.COM.

Library of Congress Cataloging-in-Publication Data

Draper, Norman Richard.  
Applied regression analysis / N.R. Draper, H. Smith. — 3rd ed.  
p. cm. — (Wiley series in probability and statistics. Texts and references section)

1. "A Wiley-Interscience publication."

Includes bibliographical references (p. — ) and index.

I. Regression analysis. I. Smith, Harry, 1923- II. Title.

III. Series.

OA278.Z.D7 1998

2.7 Durbin-Watson Test, 69	79
2.8 Reference Books for Analysis of Residuals, 70	
Appendix 2A Normal Plots, 70	
Appendix 2B MINITAB Instructions, 76	
Exercises are in "Exercises for Chapters 1-3", 96	
<b>3 Fitting Straight Lines: Special Topics</b>	
3.0 Summary and Preliminaries, 79	
3.1 Standard Error of $\bar{Y}$ , 80	
3.2 Inverse Regression (Straight Line Case), 83	
3.3 Some Practical Design of Experiment Implications of Regression, 86	
3.4 Straight Line Regression When Both Variables Are Subject to Error, 89	
Exercises for Chapters 1-3, 96	
<b>4 Regression in Matrix Terms: Straight Line Case</b>	115
4.1 Fitting a Straight Line in Matrix Terms, 115	
4.2 Singularity: What Happens in Regression to Make $X'X$ Singular? An Example, 125	
4.3 The Analysis of Variance in Matrix Terms, 127	
4.4 The Variances and Covariance of $b_0$ and $b_1$ from the Matrix Calculation, 128	
4.5 Variance of $\bar{Y}$ Using the Matrix Development, 130	
4.6 Summary of Matrix Approach to Fitting a Straight Line (Nonsingular Case), 130	
4.7 The General Regression Situation, 131	
Exercises for Chapter 4, 132	
<b>5 The General Regression Situation</b>	135
5.1 General Linear Regression, 135	
5.2 Least Squares Properties, 137	
5.3 Least Squares Properties When $\epsilon \sim N(0, I_p)$ , 140	
5.4 Confidence Intervals Versus Regions, 142	
5.5 More on Confidence Intervals Versus Regions, 143	
Appendix 5A Selected Useful Matrix Results, 147	
Exercises are in "Exercises for Chapters 5 and 6", 169	
<b>6 Extra Sums of Squares and Tests for Several Parameters</b>	149

Appendix 6A Orthogonal Columns in the $X$ Matrix, 165	
Appendix 6B Two Predictors: Sequential Sums of Squares, 167	
Exercises for Chapters 5 and 6, 169	
<b>7 Serial Correlation in the Residuals and the Durbin-Watson Test</b>	179
7.1 Serial Correlation in Residuals, 179	
7.2 The Durbin-Watson Test for a Certain Type of Serial Correlation, 181	
7.3 Examining Runs in the Time Sequence Plot of Residuals: Runs Test, 192	
Exercises for Chapter 7, 198	
<b>8 More on Checking Fitted Models</b>	205
8.1 The Hat Matrix $H$ and the Various Types of Residuals, 205	
8.2 Added Variable Plot and Partial Residuals, 209	
8.3 Detection of Influential Observations: Cook's Statistics, 210	
8.4 Other Statistics Measuring Influence, 214	
8.5 Reference Books for Analysis of Residuals, 214	
Exercises for Chapter 8, 215	
<b>9 Multiple Regression: Special Topics</b>	217
9.1 Testing a General Linear Hypothesis, 217	
9.2 Generalized Least Squares and Weighted Least Squares, 221	
9.3 An Example of Weighted Least Squares, 224	
9.4 A Numerical Example of Weighted Least Squares, 226	
9.5 Restricted Least Squares, 229	
9.6 Inverse Regression (Multiple Predictor Case), 229	
9.7 Plimar Regression When All the Variables Are Subject to Error, 231	
Appendix 9A Lagrange's Undetermined Multipliers, 231	
Exercises for Chapter 9, 233	
<b>10 Bias in Regression Estimates, and Expected Values of Mean Squares and Sums of Squares</b>	235
10.1 Bias in Regression Estimates, 235	
10.2 The Effect of Bias on the Least Squares Analysis of Variance, 238	
10.3 Finding the Expected Values of Mean Squares, 239	
10.4 Expected Value of Extra Sum of Squares, 240	
Exercises for Chapter 10, 241	

Appendix 11A How Significant Should My Regression Be?, 247  
Exercises for Chapter 11, 250

## 12 Models Containing Functions of the Predictors, Including Polynomial Models 251

12.1 More Complicated Model Functions, 251  
12.2 Worked Examples of Second-Order Surface Fitting for  $k = 3$  and  $k = 2$  Predictor Variables, 254  
12.3 Retaining Terms in Polynomial Models, 266  
Exercises for Chapter 12, 272

## 13 Transformation of the Response Variable 277

13.1 Introduction and Preliminary Remarks, 277  
13.2 Power Family of Transformations on the Response: Box-Cox Method, 280  
13.3 A Second Method for Estimation  $\lambda$ , 286  
13.4 Response Transformations: Other Interesting and Sometimes Useful Plots, 289  
13.5 Other Types of Response Transformations, 290  
13.6 Response Transformations Chosen to Stabilize Variance, 291  
Exercises for Chapter 13, 294

## 14 "Dummy" Variables 299

14.1 Dummy Variables to Separate Blocks of Data with Different Intercepts, Same Model, 299  
14.2 Interaction Terms Involving Dummy Variables, 307  
14.3 Dummy Variables for Segmented Models, 311  
Exercises for Chapter 14, 317

## 15 Selecting the "Best" Regression Equation 327

15.0 Introduction, 327  
15.1 All Possible Regressions and "Best Subset" Regression, 329  
15.2 Stepwise Regression, 335  
15.3 Backward Elimination, 339  
15.4 Significance Levels for Selection Procedures, 342  
15.5 Variations and Summary, 343  
15.6 Selection Procedures Applied to the Steam Data, 345  
Appendix 15A. Hald Data, Correlation Matrix, and All 15 Possible Regressions, 348  
Exercises for Chapter 15, 355

16.3 Centering and Scaling Regression Data, 373  
16.4 Measuring Multicollinearity, 375  
16.5 Beale's Suggestion for Detecting Multicollinearity, 376  
Appendix 16A Transforming X Matrices to Obtain Orthogonal Columns, 382  
Exercises for Chapter 16, 385

## 17 Ridge Regression 387

17.1 Introduction, 387  
17.2 Basic Form of Ridge Regression, 387  
17.3 Ridge Regression of the Hald Data, 389  
17.4 In What Circumstances Is Ridge Regression Absolutely the Correct Way to Proceed?, 391  
17.5 The Phoney Data Viewpoint, 394  
17.6 Concluding Remarks, 395  
Appendix 17A Ridge Estimates in Terms of Least Squares Estimates, 396  
Appendix 17B Mean Square Error Argument, 396  
Appendix 17C Canonical Form of Ridge Regression, 397  
Exercises for Chapter 17, 400

## 18 Generalized Linear Models (GLIM) 401

18.1 Introduction, 401  
18.2 The Exponential Family of Distributions, 402  
18.3 Fitting Generalized Linear Models (GLIM), 404  
18.4 Performing the Calculations: An Example, 406  
18.5 Further Reading, 408  
Exercises for Chapter 18, 408

## 19 Mixture Ingredients as Predictor Variables 409

19.1 Mixture Experiments: Experimental Spaces, 409  
19.2 Models for Mixture Experiments, 412  
19.3 Mixture Experiments in Restricted Regions, 416  
19.4 Example 1, 418  
19.5 Example 2, 419  
Appendix 19A Transforming  $k$  Mixture Variables to  $k - 1$  Working Variables, 422  
Exercises for Chapter 19, 425

## 20 The Geometry of Least Squares 427

20.5 Orthogonalizing in the General Regression Case, 435	
20.6 Range Spaces and Null Space of a Matrix $M$ , 437	
20.7 The Algebra and Geometry of Pure Error, 439	
Appendix 20A. Generalized Inverses $M^+$ , 441	
Exercises for Chapter 20, 444	
<b>21 More Geometry of Least Squares</b>	<b>447</b>
21.1 The Geometry of a Null Hypothesis: A Simple Example, 447	
21.2 General Case $H_0: A\beta = c$ : The Projection Algebra, 448	
21.3 Geometric Illustrations, 449	
21.4 The $F$ -Test for $H_0$ , Geometrically, 450	
21.5 The Geometry of $R^2$ , 452	
21.6 Change in $R^2$ for Models Nested Via $A\beta = 0$ , Not Involving $\beta_0$ , 452	
21.7 Multiple Regression with Two Predictor Variables as a Sequence of Straight Line Regressions, 456	
Exercises for Chapter 21, 459	
<b>22 Orthogonal Polynomials and Summary Data</b>	<b>461</b>
22.1 Introduction, 461	
22.2 Orthogonal Polynomials, 461	
22.3 Regression Analysis of Summary Data, 467	
Exercises for Chapter 22, 469	
<b>23 Multiple Regression Applied to Analysis of Variance Problems</b>	<b>473</b>
23.1 Introduction, 473	
23.2 The One-Way Classification: Standard Analysis and an Example, 474	
23.3 Regression Treatment of the One-Way Classification Example, 477	
23.4 Regression Treatment of the One-Way Classification Using the Original Model, 481	
23.5 Regression Treatment of the One-Way Classification: Independent Normal Equations, 486	
23.6 The Two-Way Classification with Equal Numbers of Observations in the Cells: An Example, 488	
23.7 Regression Treatment of the Two-Way Classification Example, 489	
23.8 The Two-Way Classification with Equal Numbers of	

23.11 Recapitulation and Comments, 499	
Exercises for Chapter 23, 500	
<b>24 An Introduction to Nonlinear Estimation</b>	<b>505</b>
24.1 Least Squares for Nonlinear Models, 505	
24.2 Estimating the Parameters of a Nonlinear System, 508	
24.3 An Example, 518	
24.4 A Note on Reparameterization of the Model, 529	
24.5 The Geometry of Linear Least Squares, 530	
24.6 The Geometry of Nonlinear Least Squares, 539	
24.7 Nonlinear Growth Models, 543	
24.8 Nonlinear Models: Other Work, 550	
24.9 References, 553	
Exercises for Chapter 24, 553	
<b>25 Robust Regression</b>	<b>567</b>
25.1 Least Absolute Deviations Regression ( $L_1$ Regression), 567	
25.2 $M$ -Estimators, 567	
25.3 Steel Employment Example, 573	
25.4 Trees Example, 575	
25.5 Least Median of Squares (LMS) Regression, 577	
25.6 Robust Regression with Ranked Residuals (rreg), 577	
25.7 Other Methods, 580	
25.8 Comments and Opinions, 580	
25.9 References, 581	
Exercises for Chapter 25, 584	
<b>26 Resampling Procedures (Bootstrapping)</b>	<b>585</b>
26.1 Resampling Procedures for Regression Models, 585	
26.2 Example: Straight Line Fit, 586	
26.3 Example: Plane Fit, Three Predictors, 588	
26.4 Reference Books, 588	
Appendix 26A. Sample MINITAB Programs to Bootstrap Residuals for a Specific Example, 589	
Appendix 26B. Sample MINITAB Programs to Bootstrap Pairs for a Specific Example, 590	
Additional Comments, 591	
Exercises for Chapter 26, 591	

## Tables

- Normal Distribution, 684  
 Percentage Points of the  $t$ -Distribution, 686  
 Percentage Points of the  $\chi^2$ -Distribution, 687  
 Percentage Points of the  $F$ -Distribution, 688

## Index of Authors Associated with Exercises

## Index

## Preface to the Third Edition

The second edition had 30 chapters; this edition has 26. On the whole (but not entirely) we have chosen to use smaller chapters, and so distinguish more between different types of material. The tabulation below shows the major relationships between second edition and third edition sections and chapters.

Material dropped consists mainly of second edition Sections 6.8 to 6.13 and 6.15, Sections 7.1 to 7.6, and Chapter 8. New to this edition are Chapters 16 on multicollinearity, 18 on generalized linear models, 19 on mixture ingredients, 20 and 21 on the geometry of least squares, 25 on robust regression, and 26 on resampling procedures. Small revisions have been made even in sections where the text is basically unchanged. Less prominence has been given to printouts, which nowadays can easily be generated due to the excellent software available, and to references and bibliography, which are now freely available (either in book or computer form) via the annual updates in *Current Index to Statistics*. References are mostly given in brief either in situ or close by, at the end of a section or chapter. Full references are in a bibliography but some references are also given in full in sections or within the text or in exercises, whenever this was felt to be the appropriate thing to do. There is no precise rule for doing this, merely the authors' predilection. Exercises have been grouped as seemed appropriate. They are intended as an expansion to the text and so most exercises have full or partial solutions; there are a very few exceptions. One hundred and one true/false questions have also been provided; all of these are in "true" form to prevent readers remembering erroneous material. Instructors can reward them to create "false" questions easily enough. Sections 24.5 and 24.6 have some duplication with work in Chapter 20, but we decided not to eliminate this because the sections contain some differences and have different emphases. Other smaller duplications occur; in general, we feel that duplication is a good feature, and so we do not avoid it.

Our viewpoint in putting this book together is that it is desirable for students of regression to work through the straight line fit case using a pocket calculator and then to proceed quickly to analyzing larger models on the computer. We are aware that



